(2)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

AD-A257 937

EPORT DATE	3. REPORT TYPE AND DATES COVERED
	Final Report 01 Sep 89-31 May 92

4. TITLE AND SUBTITLE

Development of Neural Modules Based on Si/PLST Technology for optoelectronic implementation of neural networks

5. FUNDING NUMBERS

DARPA

6. AUTHOR(S)

Professors Esener & Lee

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Univ of California
Dept of Electrical & Computer Engrg Room 007
La Jolla, CA 92093

8. PERFORMING ORGANIZATION REPORT NUMBER

92-30171

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NE
Bldg 410
Bolling AFB DC 02332-6448

Dr Craig

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFOSR-90-0018

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The objective of the research program was to design opto-electronic neuron modules communicating via free-space optical interconnects, and to develop Si/PLZT opto-electronic integrated circuit technology in order to implement these designs. First, device and system requirements for artificial neural networks were studied. Minimum performance requirements for artificial neural networks implementations were extracted. Next, two opto-electronic neural network architectures were developed. The first achieves reconfigurable optical interconnects using photorefractive crystals. This system was theoretically and experimentally investigated. The second reconfigurable weights based on Si/PLZT technology. A prototype of this system was successfully built and tested. The performance of both systems exceed the minimum requirements. The first year effort involved the design of the optoelectronic architectures, the further development of the Si/PLZT process, the development of optimal neural networks data-encoding methods, and analysis of the system performances. The remaining period entailed the experimental demonstration of the CMTM system, and

14. SUBJECT TERMS	15. NUMBER OF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

the development, characterization, and application of the D-STOP prototype system.

Final Report

for

Development of Neural Modules Based on Si/PLZT Technology for Opto-Electronic Implementations of Neural Networks

Sponsored by

Defense Advanced Research Projects Agency

Monitored by AFOSR Under Grant No. AFOSR-90-0018

Grantee

The Regents of the University of California

University of California, San Diego

La Jolla, CA 92093

Grant Period: September 1, 1989 - May 31, 1992

Principal Investigators: Sadik C. Esener (619) 534-2732

Sing H. Lee (619) 534-2413

Program Manager: Barbara Yoon

Program Monitor: Alan Craig

Table of Contents

Development of Neural Modules based on Si/PLZT Technology for Optoelectronic Implementations of Neural Networks

Sadik C. Esener, Ashok V. Krishnamoorthy, Joseph E. Ford, and Sing H. Lee

Electrical and Computer Engineering Department
University of California, San Diego
La Jolla CA 92093-0407

1. Objectives

The objective of the research program was to design opto-electronic neuron modules communicating via free-space optical interconnects, and to develop Si/PLZT opto-electronic integrated circuit technology in order to implement these designs.

First, device and system requirements for artificial neural networks were studied. Minimum performance requirements for opto-electronic neural network implementations were extracted. Next, two opto-electronic neural network architectures were developed. The first achieves reconfigurable optical interconnects using photorefractive crystals. This system was theoretically and experimentally investigated. The second architecture uses fixed, free-space, diffractive optical interconnects and electronically reconfigurable weights based on Si/PLZT technology. A prototype of this system was successfully built and tested. The performance of both systems exceed the minimum requirements.

The first year effort involved the design of the optoelectronic architectures, the further development of the Si/PLZT process, the development of optimal neural network data-encoding methods, and analysis of the system performances. The remaining period entailed the experimental demonstration of the CMTM system, and the development, characterization, and application of the D-STOP prototype system.

2. Extraction of Device and System Requirements

First, the device and system requirements for implementing biological-like neural networks were examined in order to extract the initial requirements for opto-electronic technology. These include dynamic range, fan-out/fan-in, and connectivity requirements. Theoretical investigations have shown that the fan-out or connectivity requirement of the network is problem dependent. The minimum connectivity required per neuron ($< N$) is bounded by the entropy of the problem being learned. Full connectivity is therefore preferable for general purpose implementations. Accurate fan-in is also required during the learning process for the network to converge to a valid solution. The dynamic range requirement for the interconnections has been shown to scales as $0(\text{Log } N)$, where N is the fan-out. For network sizes of 1000 fully connected neurons, both theoretical and simulation results indicate that an accuracy of 7 - 8 bits during learning is sufficient, while a lower accuracy of 1-3 bits is tolerable during operation. In certain iterative learning procedures such as back propagation, a continuous neuron transfer function is required with an accuracy of approximately 3 bits. This accuracy requirement can also be relaxed to 1 bit during operation. In summary, based on a review of theoretical and simulation results, the system requirements for a general purpose neural network implementation are given by:

- Fan-out $\leq N$
- Accurate Fan-in $\leq N$

- Weight Dynamic Range: ~7-8 bits (learning)
  ~1-3 bits (operation)
- Neuron Dynamic Range: ~1-3 bits

The potentials of opto-electronic technology versus state-of-the-art neural network implementations are shown in Figure 1. It can be seen that VLSI based neural networks do not achieve the interconnect requirements for biological-like neural networks. In addition, area considerations restrict the dynamic range, accuracy, and fan-out properties of VLSI implementations. The opto-electronic neural network implementations described below can achieve the following minimum performance:

- Number of interconnections $\geq 10^8$
- Number of interconnections/sec during operation $\geq 10^{12}$ /sec
- Number of updates/sec during learning $\geq 10^9$ /sec

3. Selection of Neuron Module Architectures

Next, two architectures were investigated for implementing opto-electronic neural networks with the minimum performance requirements stated above. The first architecture uses Si/PLZT technology to implement neuron arrays and reconfigurable volume holograms in photorefractive crystals to implement the synapses. The second architecture uses Si/PLZT technology to implement the neurons as well as their associated synapses and fixed free space diffractive optical elements for interconnections.

3.1. Reconfigurable volume holographic interconnection using photorefractives

The Correlation Matrix-Tensor Multipliers (CMTM) system described below is capable of high density interconnection between neuron arrays and has the additional advantage that the Signal-to-Noise Ratio (SNR) can be gracefully traded-in for the Space Bandwidth Product (SBP) of the optical system and components depending on the requirements of the particular neural network application.

The "standard" approach to volume holographic interconnection is to superimpose a series of Fourier-plane holograms, with at least one exposure for each input. This sequential exposure approach is impractical for large (>>100) neuron systems, because the achievable diffraction efficiency, crosstalk, and reconfiguration speed all scale poorly with the number of inputs. The CMTM is an interconnection algorithm developed in response to these problems. The algorithm is described in more detail in the attached Optic Letters manuscript. Basically, the input is transmitted through a phase-code transparency and then optically correlated with a phase-coded control image containing the connection weights (see Fig. 2). If the correlation is made using photorefractive four-wave mixing, then the interconnection can be performed and also reconfigured (updated) in parallel, avoiding the limitations of sequential approaches. Of course, this places a burden on the control image display technology. The advantage of the CMTM lies in its ability to accommodate the Spatial Light Modulator (SLM) limitations. Because the control image is held in computer memory, it is possible to manipulate it to fit the SLM size, resolution, grey-scale, nonuniformities, etc. In particular, because the input is phase-coded, it is possible to take complex grey-scaled interconnection weights, compute the ideal control image through multiplication and addition, and then truncate it to a binary phase-only image suitable for display. This reduces output SNR but does not destroy the interconnection. In general, the CMTM allows a graceful trade-off between the SNR of the interconnections and the SBP of the SLM.

This trade-off between SNR and display SBP has been investigated theoretically and using computer simulations. Both results show that the CMTM algorithm works best with dense interconnection patterns, where the number of connections to each output tends to scale with the number of inputs. When this is true, the SBP of the control image

can scale sub-linearly with the interconnection problem size (the number of weights), with little cost in SNR. This results from the statistical properties of the phase-code, using the fact that the signal adds in phase, while the noise contributions are dephased (see attached manuscript). This matches neural network interconnection requirements exactly. Initial simulations truncating the control image to binary phase only indicate that SNR is not strongly affected, although throughput efficiency is reduced.

Three experimental demonstrations were made of the CMTM algorithm. The first, which was reported in the last progress report, used degenerate four–wave mixing in 45°–cut $BaTiO_3$ to allow slowly reprogrammable interconnection (assuming the fixed input mask used in the experiment was replaced by a SLM) at the photorefractive response time. This system worked well for 9 inputs to 9 outputs, with a SNR of about 30 to 1. The second and third experiments were based on a photorefractive correlator using reflection holograms in a z–cut crystal of $LiNbO_3$. This correlator is the most accurate demonstrated to date, and was shown capable of operating on 1024x1024 pixel images. In addition, the reflection geometry allows highly selective superposition of color–multiplexed holograms.

This correlator was used to perform (potentially) fast reconfiguration between color–multiplexed interconnection holograms stored in the crystal. The system produced a SNR of about 70 to 1 for connecting 25 inputs and outputs. The SNR in the first two experiments was tested for single and multiple pixel input. The final experiment was the interconnection of large input and output arrays using a binary phase–only (compressed) control image. This system was used to test the scaling behavior of the CMTM algorithm. An aggregate SNR was measured by turning on half of the inputs simultaneously. Interconnection of up to 4096 inputs and outputs was demonstrated. The aggregate SNR for 1024 inputs and outputs was measured to be 235 to 1. The behavior followed theoretical predictions. The average SNR measured near the center of the output array was approximately half the theoretical predictions. The overall efficiency of all three experimental implementations was on the order of 1%. Some of the new experimental results using large input and output arrays are shown in Figure 3a, which shows how increasing the control image bandwidth improves the output SNR, and Figure 3b, which shows how the array size can be increased while maintaining constant control image bandwidth by trading-off output SNR.

Finally, a compact "packaged" optoelectronic processor using preprogrammed CMTM interconnection was proposed. The scaling and performance of the packaged system were computed. The result predict that a system providing one–to–one interconnection of two 1024 processor arrays with an average SNR of 10 can be contained in a volume of 40x20x10 cm^3. Alternately, a *densely* interconnected system (with 25% fill and the same SNR of 10) of two 65,536 processor arrays could fit into about a tenth this volume, 20x11x6 cm^3.

The expected performance of these CMTM systems using available display devices and crystals is well in excess of the minimum requirements in terms of high density interconnections ($>10^8$ interconnects), high speed operation ($>10^{14}$ interconnects/sec), and high speed parallel learning ($>10^9$ updates/sec).

3.2. Reconfigurable interconnection using free space optical-interconnections and electronically stored synaptic values

The second architecture that was investigated uses Si/PLZT technology to implement the neurons as well as their associated synapses and fixed free-space diffractive optical elements (CGH) to interconnect the neurons. This architecture provides full connectivity between neurons, flexible functionality neurons and synapses, accurate electronic fan-in and biologically inspired dendritic-type fan-in processing. It is

well suited to near-term implementation using VLSI technology and available SLM fabrication methods.

The opto-electronic module consists of synapses which have optical inputs and electronic outputs and reside on the same plane as their associated neurons. A typical neuron consisting of the central output unit and its surrounding synapses is shown in Figure 4. Synaptic weights are applied locally, and the results are summed via an H-tree fan-in structure. Therefore, high-accuracy summation with both positive and negative synaptic values is possible using this architecture. In addition, information processing can be performed during fan-in by placing computational units at the intermediate fan-in units of the H-tree. In Figure 5, the layout of the neuron array is shown. The detector units of one neuron are placed in the same pattern as the neurons of the array. Optical input signals from other neurons are imaged onto individual detector units via a space-invariant optical interconnection system. This dual-scale invariant layout has the property that full connectivity can be achieved optically achieved optically using demagnification and replication (Figures 6 and 7) which can be performed using a single, fixed, space-invariant Computer Generated Hologram (CGH).The number of modulators needed by such a system (N) is much lower than that required by traditional Matrix-Vector architectures (N). In addition, this system can perform both matrix-vector and outer-product operations which are needed to implement back propagation learning. Since the electronic line lengths of the H-tree are equal, there is no timing skew between signals arriving at the central unit from different synapses. This feature can be used to preserve precise timing relationships between firing neurons (Asynchronous mode) or can be used to pipeline operations and thereby increase the speed of operation (Synchronous mode)

The storage capacity of this system can be enhanced when used in conjunction with a secondary storage device such as the UCSD parallel readout optical disk. Near-term implementations of this system will easily meet the minimum storage requirements ($>10^8$ interconnects using optical disk), operation speeds ($>10^{12}$ interconnects/sec), and learning rates ($>10^9$ updates/sec). As discussed in section 6, a system prototype was built and tested.

4. Further Development of Si/PLZT Technology

The research effort was subdivided into two main areas: i) material development to enhance Si/PLZT technology and ii) device development. Major emphasis on the material development was put on the magnetron sputter thin PLZT films on silicon on sapphire. On device development, initial research effort was concentrated in the design and fabrication of 1-D SLM in laser recrystallized poly-silicon on PLZT. We also designed the flip-chip bonded Si/PLZT SLMs suitable for the D-STOP system.

4.1. Magnetron Sputtering of Thin PLZT Films on Silicon-on-Sapphire

Lanthanum modified lead zirconate titanate (PLZT) is a transparent ferroelectric material exhibiting large electro-optic effect. In particular, PLZT (9/65/35) in perovskite phase has the largest electro-optic coefficient in the family. The deposition of PLZT (9/65/35) on r-plane sapphire may enable the integration of electro-optic devices with silicon circuits, since epitaxial silicon can be successfully grown on (1102) r-plane sapphire.

There are some fundamental challenges in the deposition of PLZT (9/65/35) films. PLZT films tend to exist in pyrochloric phase when the substrate temperature during deposition is less than 450°C. To obtain perovskite PLZT films the substrate temperatures should be high (>550°C). However high temperatures result in excessive loss of lead.

To meet the challenges outlined above, the following modifications were implemented:

i) Compensation of lead by lowering deposition temperature and by using one PLZT source.
ii) Post-deposition annealing at higher temperatures to improve the existence of perovskite phase in the film.
iii) Low-temperature deposited oxide as a capping layer to prevent the out-diffusion of lead during annealing.

A triode magnetron sputtering system was used in our deposition experiments. The triode sputtering system can operate at low chamber pressures (< 1 Pa), typically an order of magnitude less than conventional planar magnetron sputtering systems. The triode sputterer is extremely versatile in its ability to obtain controlled, reproducible plasma conditions and thus is suited to achieve stable, high quality PLZT films. It uses two R-F triode magnetron sputter sources and one Ion gun, with low energy Ar ions to permit pre-deposition substrate cleaning, in a planetary configuration. The sapphire substrate is epitaxially polished on the front face (the face of film growth) and optically polished on the back face and mounted on a rotating Molybdenum backing plate using molten indium and heated on the back side using radiant heating by means of halogen lamps. The substrate temperature is monitored using a Chromel-Alumel thermocouple attached to the substrate holder.

The films were deposited on a rotating substrate and satisfactory results were obtained on the uniformity of film thickness and composition over the whole area of the substrate. The target composition was determined so that the right stoichiometry of the deposited film could be obtained at temperatures of around 500°C. The sputtering conditions used in the experiment are summarized in Table 1. Post deposition anneal in an oxygen atmosphere was performed at temperatures between 600°C and 800°C to improve the properties of the deposited film. A layer of 200nm thick Silicon dioxide was deposited by Plasma Enhanced Chemical Vapor Deposition (PECVD), prior to introduction of the sample into the furnace, prevent loss of lead from the film. The annealing temperature and time was adjusted to maintain the film stoichiometry, while at the same time improving the film properties.

Target Diameter	0.05m
Target to substrate spacing	0.250m
R.F. Power	100-160
W Sputtering Gas	Argon
Chamber Pressure	0.5 Pa
Substrate Temperature	400 - 600°C
Growth rate	2 -3 nm/min

Table 1. Sputter Deposition Conditions

Compositional Analysis of the film were performed by Rutherford Backscattering Spectrometry (RBS) using a 2.3 Mev He^{++} beam. Analysis on bulk PLZT samples of known composition were performed to establish a baseline for comparison. Thin film samples both as deposited and after post-deposition anneal were analyzed and the experimentally obtained spectra was super imposed over the simulated spectra for the case of PLZT (9/65/35) on sapphire.

A plot of the film composition as a function of deposition temperature is depicted on Figure 10. The investigation was carried out in the range of 350°-600°C with a single

target PLZT source of an exact composition of 9/65/35. As can be seen there was a rapid loss of lead from the film when the deposition temperature was raised beyond 550°C. Lead depletion was observed at lower deposition temperatures than expected. This is probably due to the geometry of the system. The distance between the target and the substrate plane is large resulting in low deposition rate in spite of a large sputtering rate on the target (obtained by applying a high R-F power to the targets). The large source to substrate distance is, however, required to insure good uniformity over the whole area of the substrate. By holding the substrate temperature at 500°C and adjusting the composition of the PLZT target, we were able to accurately control the composition of the PLZT film. The results are shown in Figure 11. R-F power on PLZT target was 160 Watts. The composition of the film is evaluated from RBS spectra as follows. The step heights are proportional to the ratio of the element in the film. The step widths are proportional to the thickness of the film. By carefully locating the step positions it is possible to identify the presence and the amount of individual elements in the film. Furthermore the background obtained from the substrate serves as a basis to obtain the accurate composition of the film. It can be shown that the composition of the film is indeed PLZT (9/65/35). Auger Electron Spectroscopy (AES) was also used to correlate the results of RBS and obtain accurate information about the compositional uniformity of the film as a function of depth from the surface.

Crystallographic studies were done using X-ray diffraction with CuK_α radiation. As can be seen from the X-ray diffraction data shown in Fig. 12 the as-grown film had a good crystalline structure with a dominant (110) orientation. However a pyrochloric peak at 34° is also seen. The pyrochloric peak could be due to lead deficiency. In general the pyrochloric films are yellowish in color and the perovskite films colorless. In our experiments, lead compensation changed the color of the film, from yellowish to colorless, indicating a change in crystal structure. We observed, in crystallographic studies, that annealing suppressed the pyrochloric peaks. SEM studies indicated featureless surfaces with a smooth texture which would imply a good degree of epitaxial growth. The results show that we were successful in depositing PLZT 9/65/35 films of right stochiometry.

Film thicknesses were in the range of 0.3-0.8 μm and they were initially estimated by RBS. Ellipsometry results were used to evaluate both the refractive index and thickness of the films. There was a good correlation between the thickness as obtained through RBS and ellipsometry. The refractive indices of the films, measured at 632 nm was in the range of 2.2-2.5. This is close to the reported value of 2.5 of bulk PLZT.

In order to measure the electrical and electro-optic properties of the films, aluminum was deposited by thermal evaporation and patterned using standard photolithography to form an interdigitated electrode structure. The dielectric constant of the perovskite films were measured in the range of 1500-2500 (at 10Khz) whereas the as-deposited pyrochloric films had a dielectric constant in the range 10-100. The dielectric constant tend to decrease at higher frequencies (similar effect was observed in our measurements on bulk samples). A plot of the dielectric constant as a function of temperature for frequencies of 10Khz, 100Khz and 1Mhz is shown in Fig. 13. As is evident, the dielectric constant shows an anomaly at about 220°C. The curie temperature thus obtained is larger than the reported value for bulk samples. This could be due to microscopic changes in the crystal structure of the films as compared to the bulk ceramics.

To measure the electro-optic coefficient of the films, coherent light at 514 nm was polarized 45° with respect to the electric field applied on the sample and was incident on the film. The output intensity from a crossed analyzer was detected. This intensity is proportional to the change in refractive index δn caused by the applied field:

$$\delta n = -\frac{1}{2} n^3 R E^2$$

where R is the quadratic electro-optic coefficient of the film, n is the refractive index of the film and E is the electric field applied on the film. A plot of change in the index of refraction as a function of applied electric field is shown in Fig. 14. The electro-optic coefficient was calculated to be 0.6×10^{-16} m^2/V^2.

We are currently working for better understanding of the electro-optics properties of the deposited PLZT films and for the fabrication of light modulators with deposited films.

4.2. Design of Hybrid Flip-chip bonded Si/PLZT SLMs

A promising approach toward the integration of silicon circuits and PLZT modulators is use a hybrid integration technique, such as flip-chip bonding. Flip-chip bonding, currently used for silicon packaging, is a mature and well developed technique that can also be used advantageously to realize S-SLMs. This technique has been studied with different materials and devices for hybrid conventional SLMs. Figure 15 illustrates this approach, and shows the design of a neuronal element in a hybrid silicon/PLZT S-SLM. The PLZT wafer is used to support the PLZT modulators. The modulator is used in a reflective configuration and is connected electrically using flip-chip bonding to the output of the silicon circuit in the PE with an indium bump. The size of the bump which can be made as small as 10 μm, is governed by the warpage of the wafers and the desired distance between wafers. This approach is particularly promising for the D-STOP neural system, since the density of modulators is low. It is expected that 32 by 32 arrays of optoelectronic neurons with their associated synapses (hence a large neuron complexity) can be readily produced with this hybrid integration technique.

5. Development of Data Encoding Methods/Si Circuitry

Next, the question of data representation was considered. An important issue in the implementation of an OE neural system is the choice of data encoding methods. For the optical neuron-to-synapse channel, pulse-amplitude-modulation (PAM) schemes which require high contrast ratio light modulators and tightly controlled optical interconnects impose serious system constraints. Among the binary encoding schemes, pulse-frequency-modulation (PFM) demands high dynamic power since the light modulator and detector capacitances must be charged and discharged at high speeds. A binary pulse-width-modulation (PAM) scheme however, is highly energy efficient since only two state transitions per communication is needed. For the local electronic synapse-to-neuron channel on the other hand, Available VLSI devices and circuit techniques can provide the required precision for PAM methods to be used with high integration densities. The use of a different modulation scheme in the synapse also increases the dynamic range of the synaptic multiplication.

For these reasons, a PWM method for optical neuron-to-synapse communication and a PAM scheme for the electronic synapse-to-neuron communication are well suited for an OE neural system implementation. In this encoding method, neurons modulate the width of aperiodic, clocked, pulse stream. The width or duration of this optical pulse contains the information (<3 bits) of the neuron-output. This binary amplitude encoding method allows reliable, low energy communication even after passing through an optical interconnect with a low SNR and high fan-out. The width of the received pulse is then multiplied by the synaptic value (in time) and then integrated to find the total neuron input. Since this multiplication is done in time, no complicated decoding circuitry is

needed. In addition, digital storage combined with simple D/A converters can provide accurate synaptic values without requiring large-area digital multipliers.

6. Implementation of Neural System Prototype

6.1. Description of the Prototype Neural System

The prototype system consists of a 16-node input, 4-neuron hidden and a single neuron output layer. The picture and the network equivalent of the optoelectronic neural network system is shown in Figure 16.

A lenslet array is used to focus the filtered, collimated and polarized argon laser beam on a 4 by 4 array of electrically addressed PLZT light modulators which constitute the input (distribution) layer to the network. Two lenses demagnify the image of the light modulators to the size of one neuron. A polarizing beam splitter between these lenses analyses the output of the light modulators and splits the image for visual inspection of the applied input. A hologram in conjunction with a third lens replicates the demagnified 4 by 4 image of the light modulators over the 4 hidden layer neurons in the OE neural chip to achieve full connectivity. An additional board which is electrically connected to the neural chip, houses the output neuron together with its 4 synapses. 5 LED's on this board display the outputs of hidden and output layer neurons and 4 others show the analog outputs of the 4 detectors at each corner of the chip to enable the alignment. A digital tester (LV500) generates system clocks and downloads the synaptic weights (learned off-system) to the neural chip. In the following sections, we give detailed information on the key system components.

6.2. Input PLZT Modulator Array

Figure 17 shows the cross-section of a typical electrically-addressed PLZT light modulator. An oxide layer was deposited on bulk PLZT, patterned and metallized. Electrode spacing was 15 microns. A contrast ratio of 20:1 (among the array) was measured. The average modulated optical power was 200μW per modulator. Up to several tens of megahertz switching speeds is possible using appropriate drivers.

6.3. D-STOP Optical System

The D-STOP optical system is shown in figure 18. The first two lenses form a demagnified image of the input array of modulators. The third lens transfers this image to the output plane. A holographic beamsplitter in contact with the third lens performs the replication and can also provide aberration correction. Because the light shares a common path, there is no small aperture bottleneck, and the diffraction-limited resolution of the system is high. The telecentric demagnifying stage maintains high throughput efficiency, and separates the holographic beamsplitter from the short focal length demagnifying lens, allowing a fixed maximum diffraction angle. The current prototype system uses 4 by 4 input and 8 by 8 output array. Experimental results related to larger 8x8 input and 64x64 output arrays were reported earlier.

6.4. Optoelectronic Neural Chip

The layout of the optoelectronic neural chip which implements the hidden layer of the neural system is shown in figure 19. In each neuron the outputs of 4 synapses are connected to a fan-in unit and the outputs of 4 fan-in units are connected to the neuron

soma. The sub-units of the optoelectronic neuron are discussed in the following subsections.

6.4.1. The synapse unit :

The functional block diagram of the optoelectronic synapse unit is shown in figure 20.

Each synapse receives the PW-modulated optical input through a light detector at its center, the output of which is restored to control a 5-bit current scaling, multiplying D/A converter (MDAC)[9]: as long as the input optical pulse is high (i.e. synapse is illuminated) , the MDAC converts the digitally stored synaptic weight into an analog current which, depending on the sign of the weight, is sourced to either the excitatory or the inhibitory output of the synapse.

The use of pulse-width modulation to encode the neural outputs eases the synapse implementation: the synapse simply modulates the amplitude of the pulse-width modulated presynaptic inputs. An integration of the synaptic output signal, which is analog both in amplitude and pulse-width , therefore provides the result of the analog synaptic multiplication.

Figure 21 shows the photodetector circuitry of the synapse unit.

A reverse-biased p-n junction is used to generate an electronic photocurrent from the optical input. PMOS load transistor converts the photocurrent into a photovoltage which is then thresholded by a CMOS inverter to generate an electronic bit. Variable gate voltage (VX) of the load transistor allows the control of detectivity. The well-to-substrate p-n junction (in a typical MOSIS p-well process) was used to build the photodiode : active diode area was 15μm by 15μm. The responsivity was measured to be 0.33 A/W. 10 μW optical power was needed to detect the light at a speed of 1MHz.

Figure 22 shows the synapse output current as a function of the digitally stored synaptic weight. High linearity of the D/A conversion is observed.

6.4.2. The fan-in unit :

The dual output (excitatory and inhibitory) channels of a group of four synapses are connected to the analog fan-in (dendrite) unit (figure 19) where the combined inhibitory output currents of the four synapses are subtracted from their total excitatory output current. The difference current is also demagnified to ease the summation of currents at the neuron soma.

The use of differencing fan-in units removes the need for bipolar synapses requiring both current sourcing and sinking capabilities. This allows the use of unipolar synapses offering smaller area and well-matched synaptic output currents.

Figure 23 shows the functional block diagram of the fan-in unit.

Two current mirrors, one with a reversed output current direction implements the subtraction operation. A scaling factor a is also introduced during current mirroring. The regulated-cascode current mirror stabilizes the output current by increasing the output resistance of the unit. This way, the fan-in unit also acts as a buffer between the neuron soma and its synapses, thereby reducing any undesirable negative feedback from the neuron soma to its synapses. Figure 24 shows the fan-in unit output current as a function of the difference in total excitatory and inhibitory current inputs to the fan-in unit. Dendrite unit preserves the linearity of the fan-in processing.

6.4.3. The neuron soma unit :

The analog, bipolar outputs of four fan-in units are integrated at the neuron soma to give the net input activity voltage to the neuron. This voltage is then converted into a neural output pulse, the width of which represents the output information of the neuron.

Figure 25 shows the functional block diagram of the neuron soma which consists of a spatio-temporal integrator, a voltage-to-pulse-width converter and buffers to drive the neuron output capacitance.

The integrator consists of a simple clocked capacitor : during the time window defined by two non-overlapping clock phases (f1 and f2), the capacitor integrates the fan-in unit outputs (I_{f1}-I_{f2}) . The final value of the voltage integrated on this capacitor represents the net input activity to the neuron. An RC circuit converts this voltage into a pulse width : the greater the integrated input activity voltage, the wider the output pulse. The neuron gain is introduced by controlling the discharge rate of the RC circuit. Figure 26 shows the width of the neural output pulses as a function of the integrated input activity for different gain control (V_H) parameters.

A minimum neural output pulse of approximately 100nsec. was measured. This suggests a maximum network speed of approximately 640 million interconnections per second.

The chip was implemented in a 2μm, double-poly p-well process through MOSIS. The chip area is 4mm by 6mm. Average power consumption was 20mW and thus negligible.

6.5. Optoelectronic Neural System Application

We ran a simple recognition application to test the system. A neural network which matches the prototype system size was trained on a SUN computer to distinguish the vertical lines from the horizontal ones in a 4 by 4 pixel image. The application was trained using noisy horizontal and vertical lines. The determined synaptic weights were then downloaded to the OE neural chip to test the hardware. Figure 27 shows the patterns applied to the system together with the system's response.

The network was capable of distinguishing any of the four horizontal lines from any of the four vertical ones shown above.

7. Conclusions

The objective of the research was to design opto-electronic neuron modules communicating via free-space optical interconnects, and to develop Si/PLZT opto-electronic integrated circuit technology in order to implement these designs.

During the course of this research program, we have demonstrated how optoelectronic technology can offer solutions to interconnection problems in the implementation of neural networks. We began by considering hardware requirements imposed by large-scale neural network paradigms. We then examined the key optoelectronic system components including S-SLMs, free-space optical interconnects and parallel accessed optical storage devices, and discussed and analyzed their potential use for scalable neural systems. We further developed and improved the Si/PLZT process. We then provided two specific examples of how the use of optoelectronic technology can lead to high performance, near-term solutions to scalable neural systems. We designed and experimentally verified a volume holographic technique (CMTM) that

can be used for large-scale neural networks. We also designed a general 3-D opto-electronic neural system (D-STOP) that uses a hybrid silicon VLSI based opto-electronic integrated circuit technology to implement the neurons and their associated synapses, and fixed, free-space optical diffractive elements to interconnect the neurons. The D-STOP architecture provides full connectivity between neurons, flexible functionality for neurons and synapses, accurate electronic fan-in, biologically-inspired dendritic-type fan-in processing, and minimizes the number of required light transmitters. Finally a fully operational D-STOP system prototype was built and tested. The system was characterized and applied to a simple pattern recognition problem.

8. List of Publications

1. G. Marsden, A. Krishnamoorthy, S. Esener, And S. H. Lee "Dual-Scale Topology Optoelectronic Processor," Paper TuJJ2, *Proc. OSA Annual Meeting* Boston 1990, p. 109.

2. A. Krishnamoorthy, J. Ford, G. Marsden, G. Yayla, S. Esener, and S. H. Lee, "D-STOP: Comparative Analysis and Technological Feasibility," *Proc. OSA Topical Meeting on Optical Computing*, Salt Lake City, p. 244, March 1991.

3. G. Marsden, B. Olsen, S. Esener, and S. H. Lee, "Optoelectronic Fuzzy Logic \System," *Proc. OSA Topical Meeting on Optical Computing*, Salt Lake City, p. 212, March 1991.

4. A. Krishnamoorthy, G. Yayla, and S. Esener, "Design of a Scalable Optoelectronic Neural System using Free-Space Optical Interconnects," *Proc. Intl. Joint Conf. on neural Networks*, p. 527, July 1991.

5. A. Krishnamoorthy, G. Marsden, G. Yayla, J. Ford, and S. Esener, "Dual-Scale Topology Optoelectronic Processor," *UCSD Technical Report*, July 30, 1991.

6. G. Marsden, A. Krishnamoorthy, S. Esener, and S. H. Lee "Dual-Scale Topology Optoelectronic Processor," *Optics Letters*, Vol. 16, No. 24, December 15, 1991.

7. G. Marsden, A. Krishnamoorthy, J. Mercklé, G. Yayla, J. Ford, and S. Esener, "D-STOP: An Optoelectronic Generalized Matrix Algebra Processor," *Proc. CAMP-91*, International Conf. on Computer Architecture for Machine Perception, December 1991, Paris France.

8. A. Krishnamoorthy, G. Yayla, and S. Esener, "Scalable Optoelectronic Neural System using Free-Space Optical Interconnects," *IEEE Transactions on Neural Networks*, Vol. 3, No. 3, pp. 404-413, May 1992.

9. G. Yayla, A. Krishnamoorthy, G. Marsden, J. Ford, G. Marsden, V. Ozguz, C. Fan, S. Krishnakumar, and S. Esener, "Prototype 3-D Optoelectronic Neural System," *Proc. SPIE Annual Meeting*, San Diego, July 1992.

10. A. Krishnamoorthy, G. Yayla, G. Marsden, and S. Esener, "Free-Space Optoelectronic Neural System Prototype," *to be presented at OSA Annual Meeting*, Alburquerque NM

11. S. Krishnakumar, S. Esener, C. Fan, V. Ozguz, M. Title, C. Cozzolino, and S. H. Lee "Characterization of Ferroelectric Thin Film PLZT(9/65/35) on R-Plane Sapphire," *Proc. MRS*, April 1990.

12. S. Krishnakumar, V. Ozguz, C. Fan, M. Title, C. Cozzolino, S. Esener, and S. H. Lee "Deposition and Characterization of Thin Ferroelectric PLZT Films for Spatial Light Modulator Applications," *IEEE Trans. on Ultrasonics, Ferroelectrics, and Frequency Control*, Vol. 38, No. 6, Nov. 1991.

13. J. E. Ford, Y. Fainman, and S. H. Lee, "Array interconnection by phase coded optical correlation," *Opt. Lett.* **15**, 1088-1090, 1990.

14. J. E. Ford, "Reconfigurable Array Interconnection by Photorefractive Volume Holography," *Ph.D. Dissertation*, Chp. 4, 1991.

9. APPENDIX

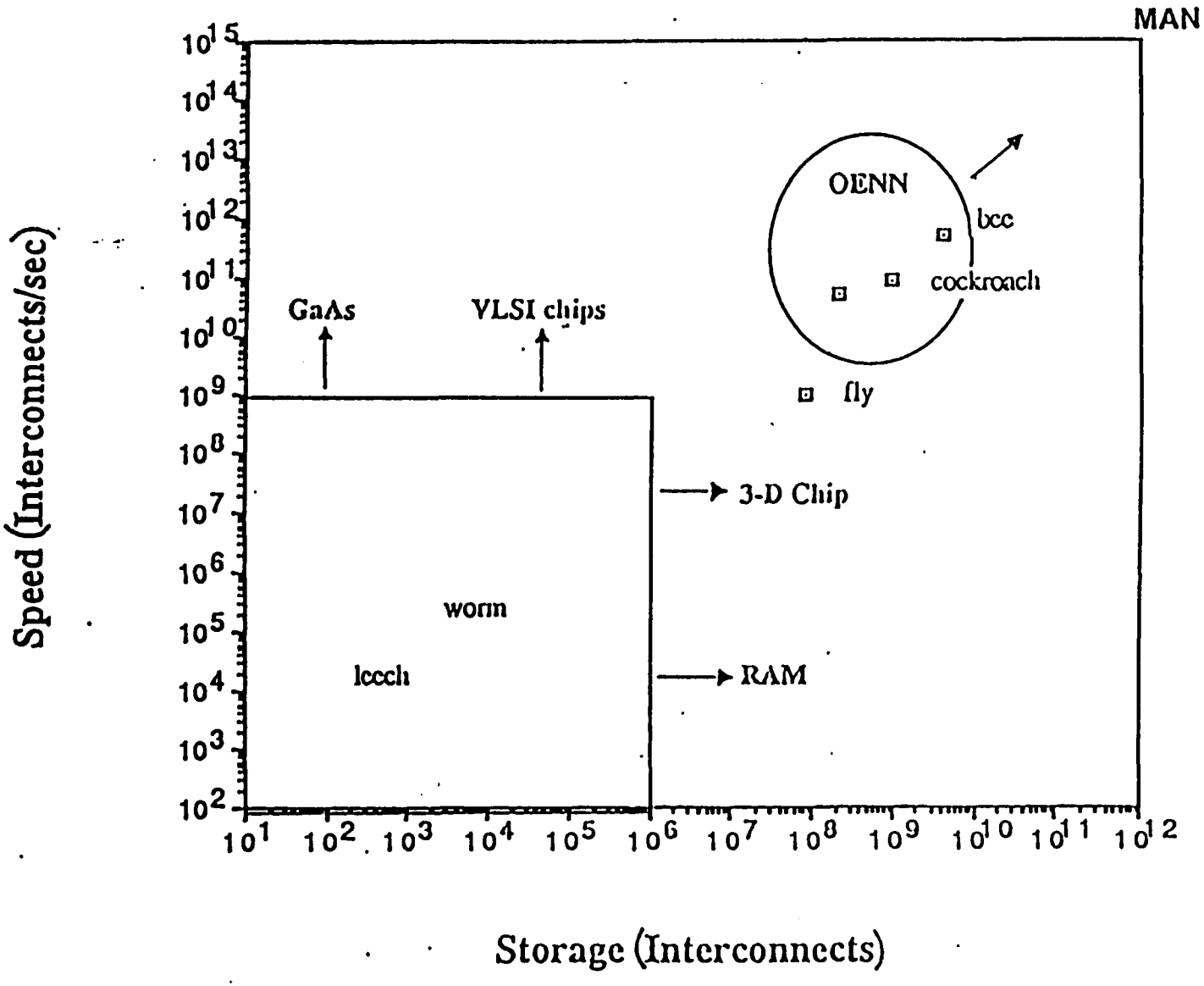


Figure 1. Potentials of Optoelectronic Neural Networks

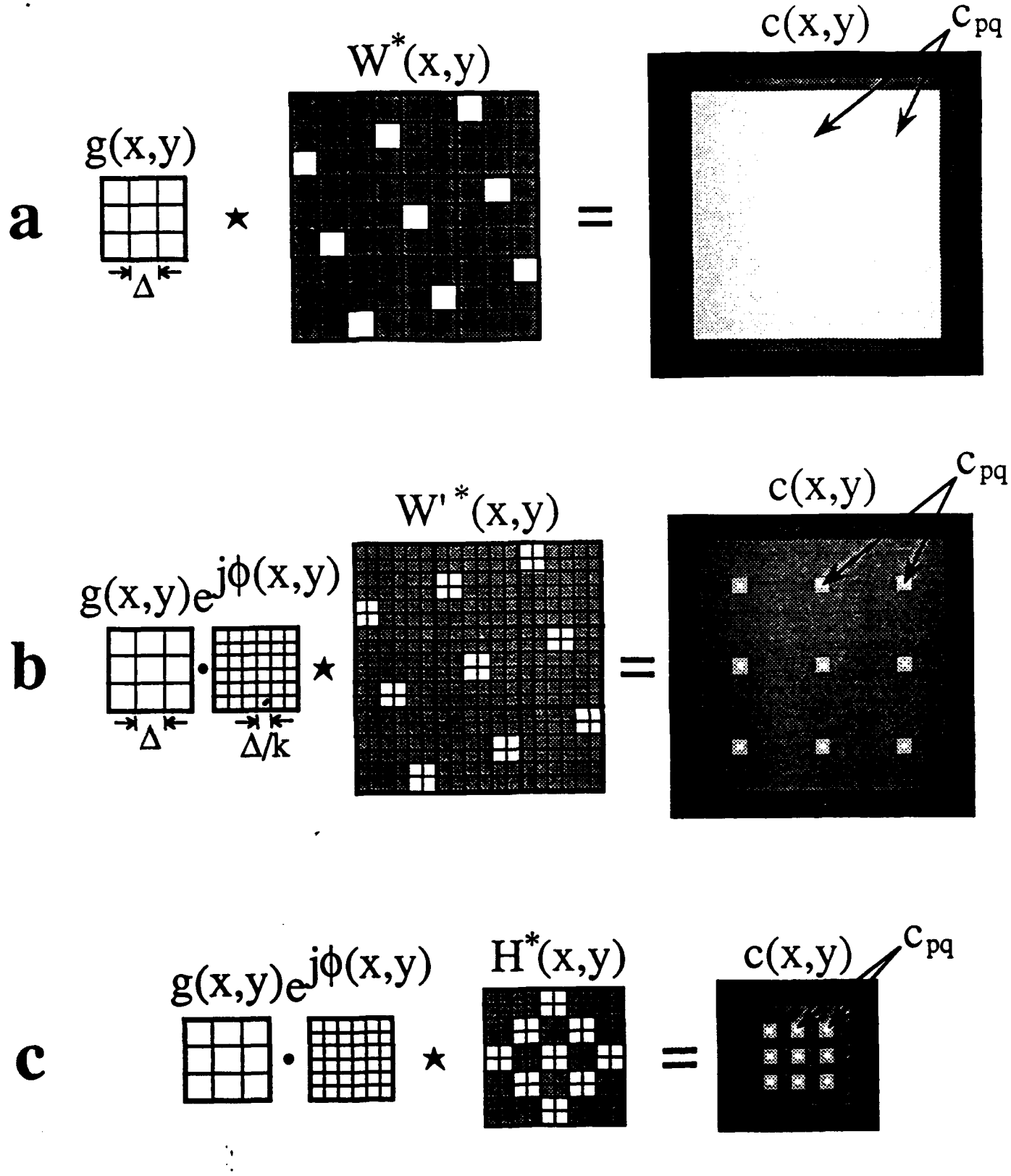


Figure 2. Conceptual construction of the CMTM algorithm in three steps. 1a: Correlation of the input array with the control image produces the connected output at sites imbedded in a field of noise. 1b: Phase-coding the input and control images reduces background noise relative to the signal. 1c: The final control image $H(x,y)$ is produced by compressing (overlapping) $W'(x,y)$, reducing control SBP at the cost of superimposing noise on the signal sites.

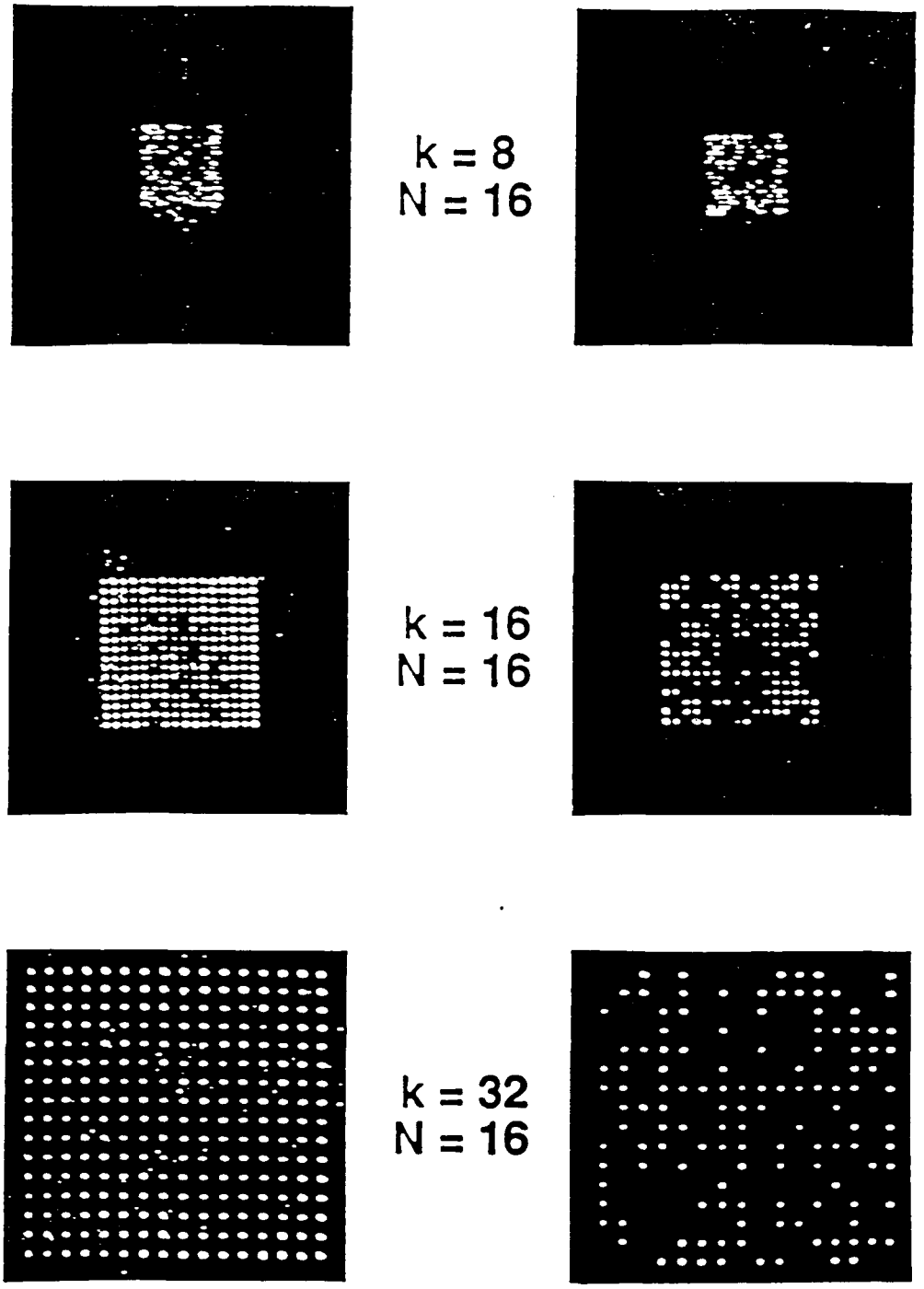


Figure 3a : Effect of phasecode density **on output** SNR. Phasecode pixel size **was constant** at 20 μm, so as k increases **the array size also** increases. At left, all inputs **on. At right,** half the inputs masked. Signal **intensity and** uniformity increase as k grows.

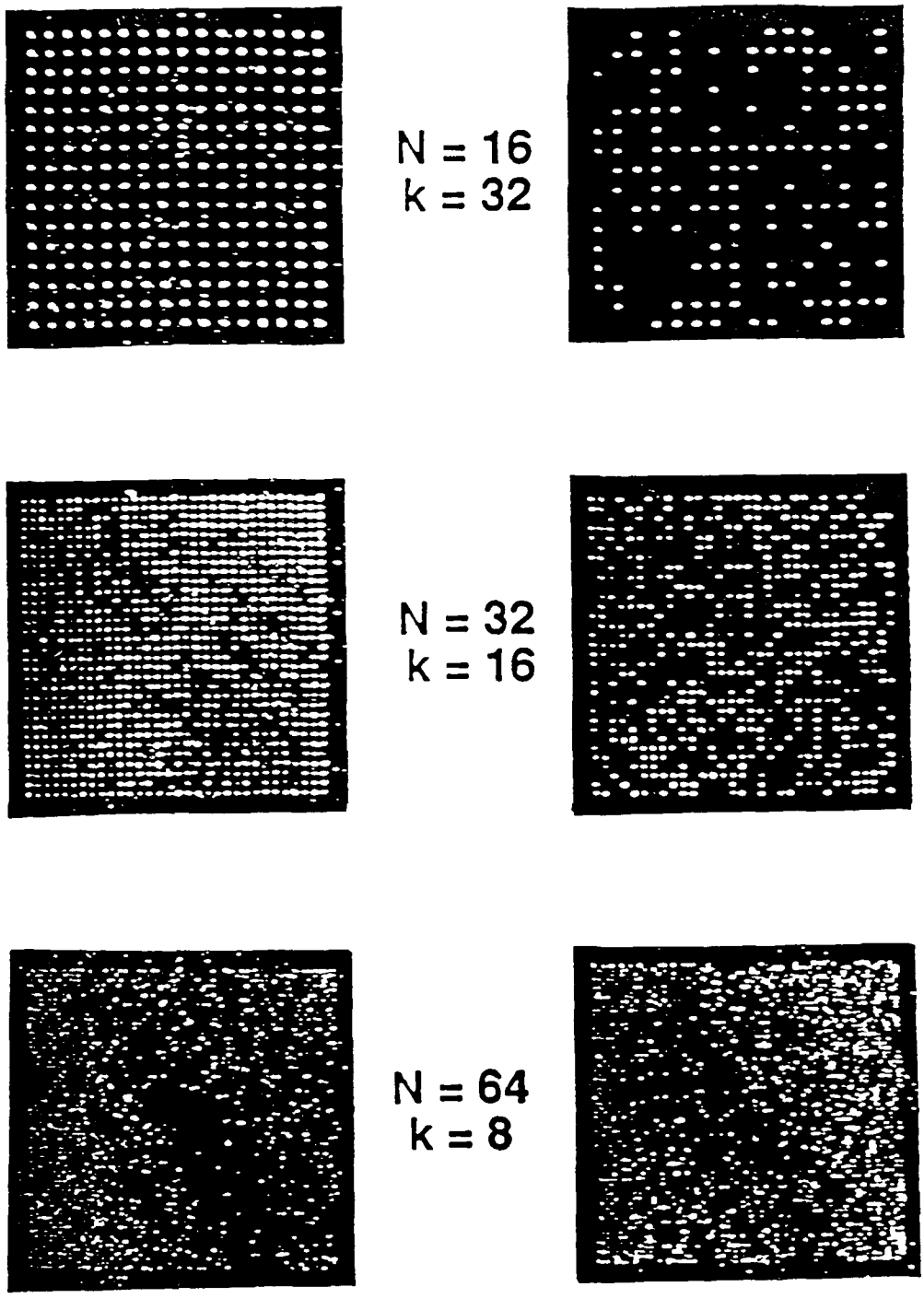


Figure 3b : Effect of array size on output **SNR. The** control image SBP was held **at about** 1000x1000 by decreasing k linearly **with N. At left**, all inputs on. At right, **half the inputs** masked. Output SNR decreases **as N^2 grows** from 256 to 1024 to 4096.

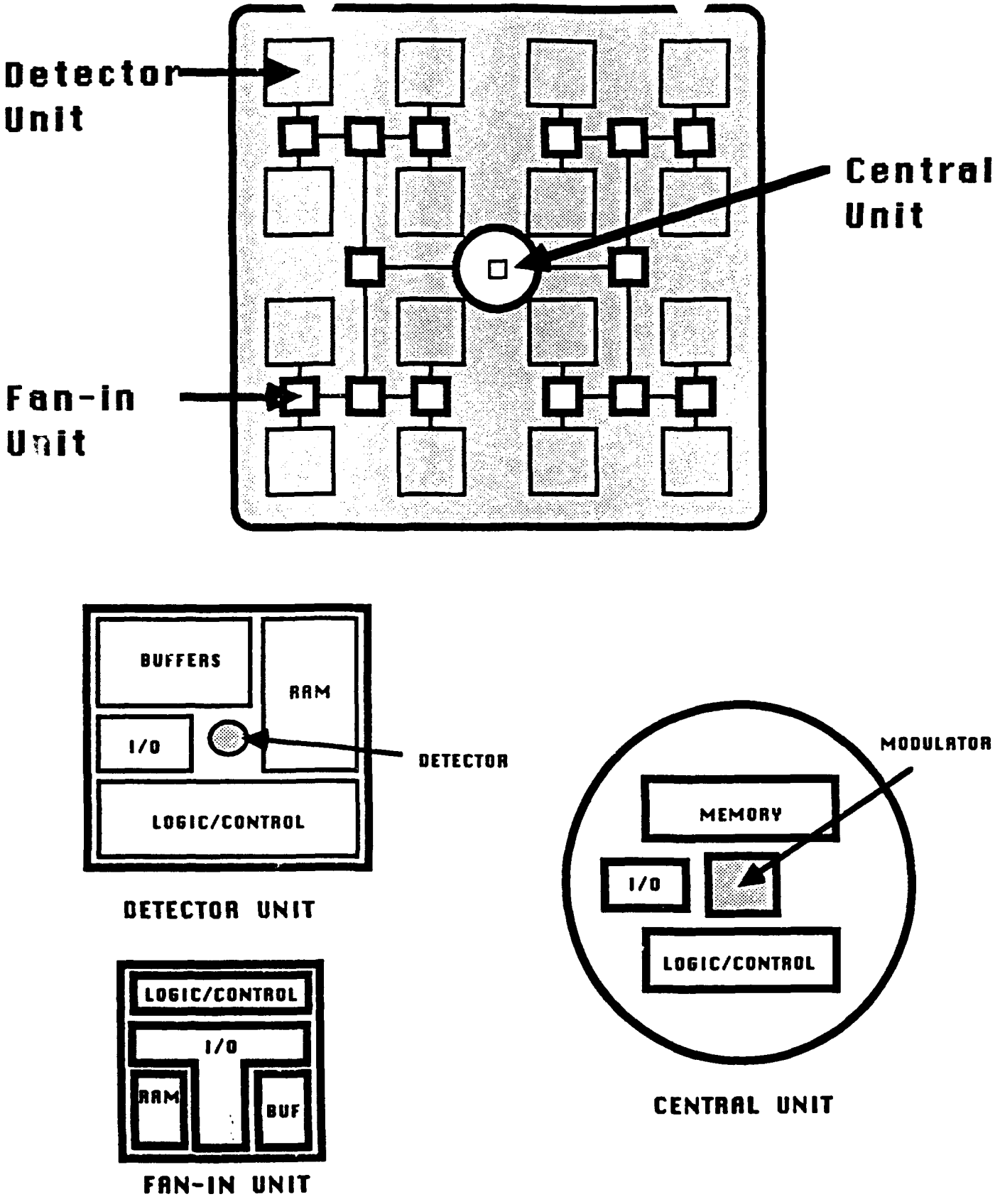


Fig 4: Neuron layout showing detector (input) units, fan-in units, and central (output) unit

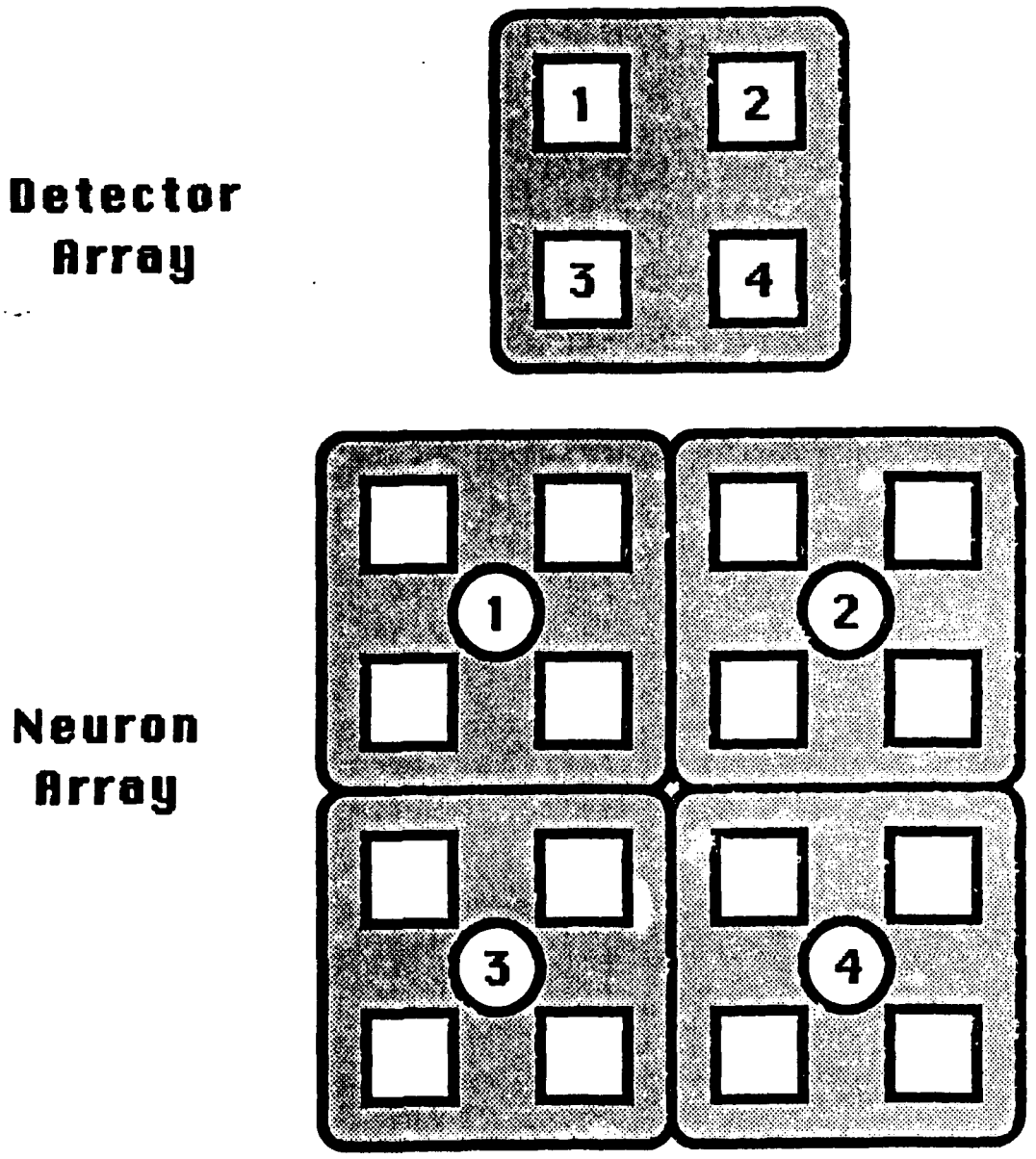


Fig 5: Opto-electronic chip layout showing dual-scale invariant neuron and detector arrays: Modulators M1-M4 have the same relative placement as the detectors D1-D4 of any one neuron

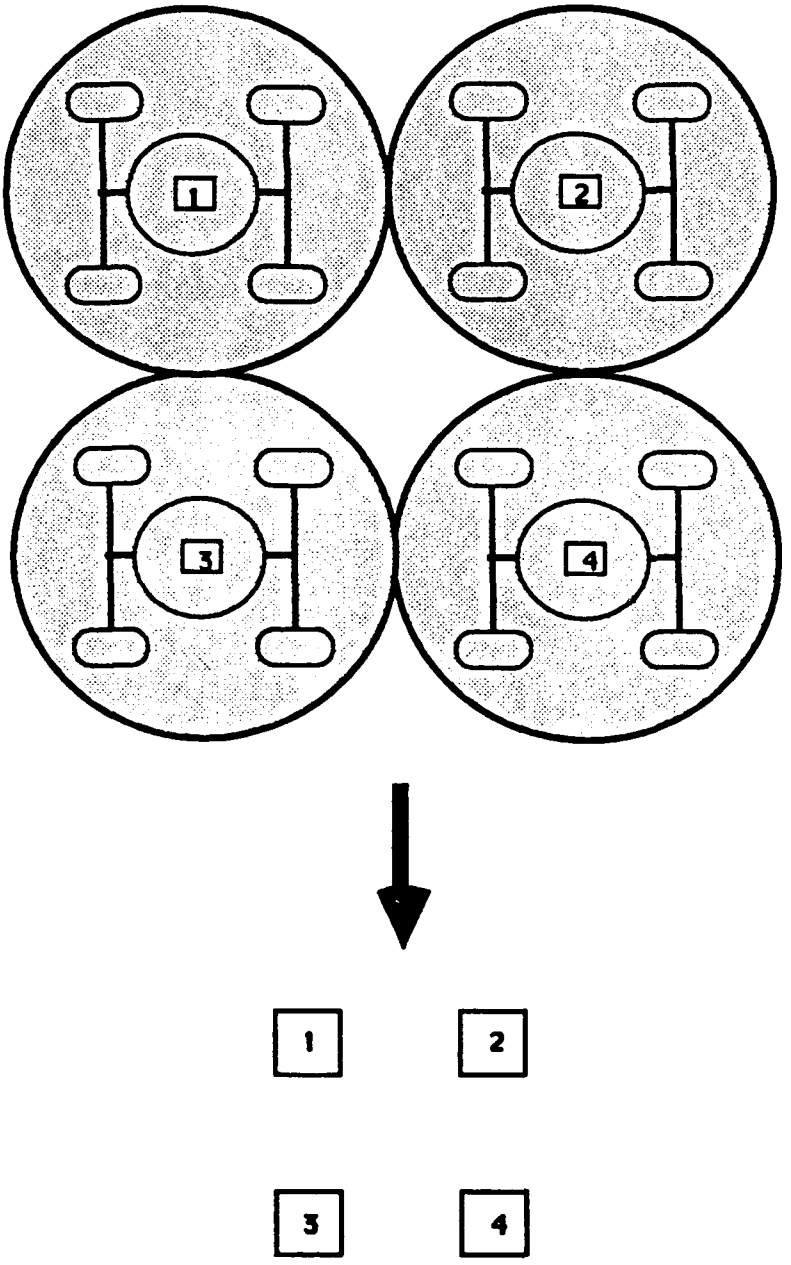


Fig 6: BROADCAST - Demagnification; The image of the modulators is demagnified to the scale of the detector units of a single neuron

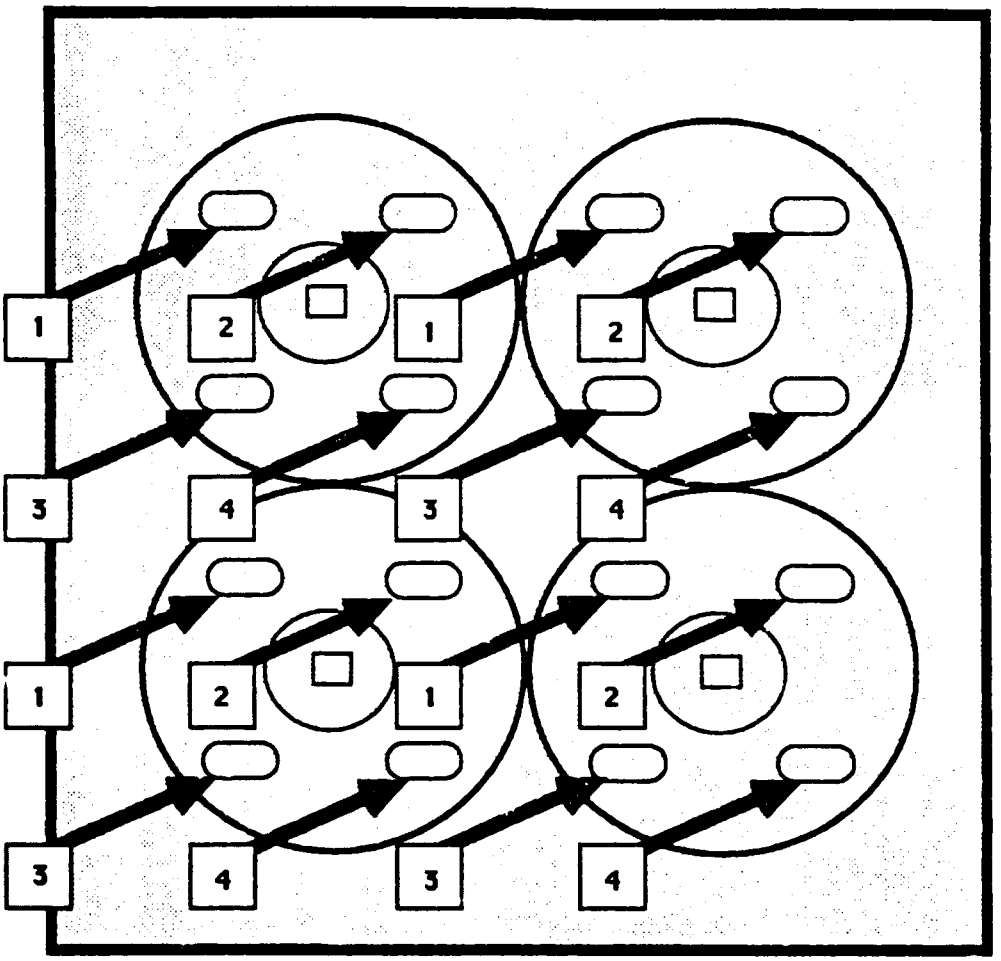


Fig 7: BROADCAST-Replication: The demagnified image of the modulators is replicated over the entire array to achieve full interconnection between neurons

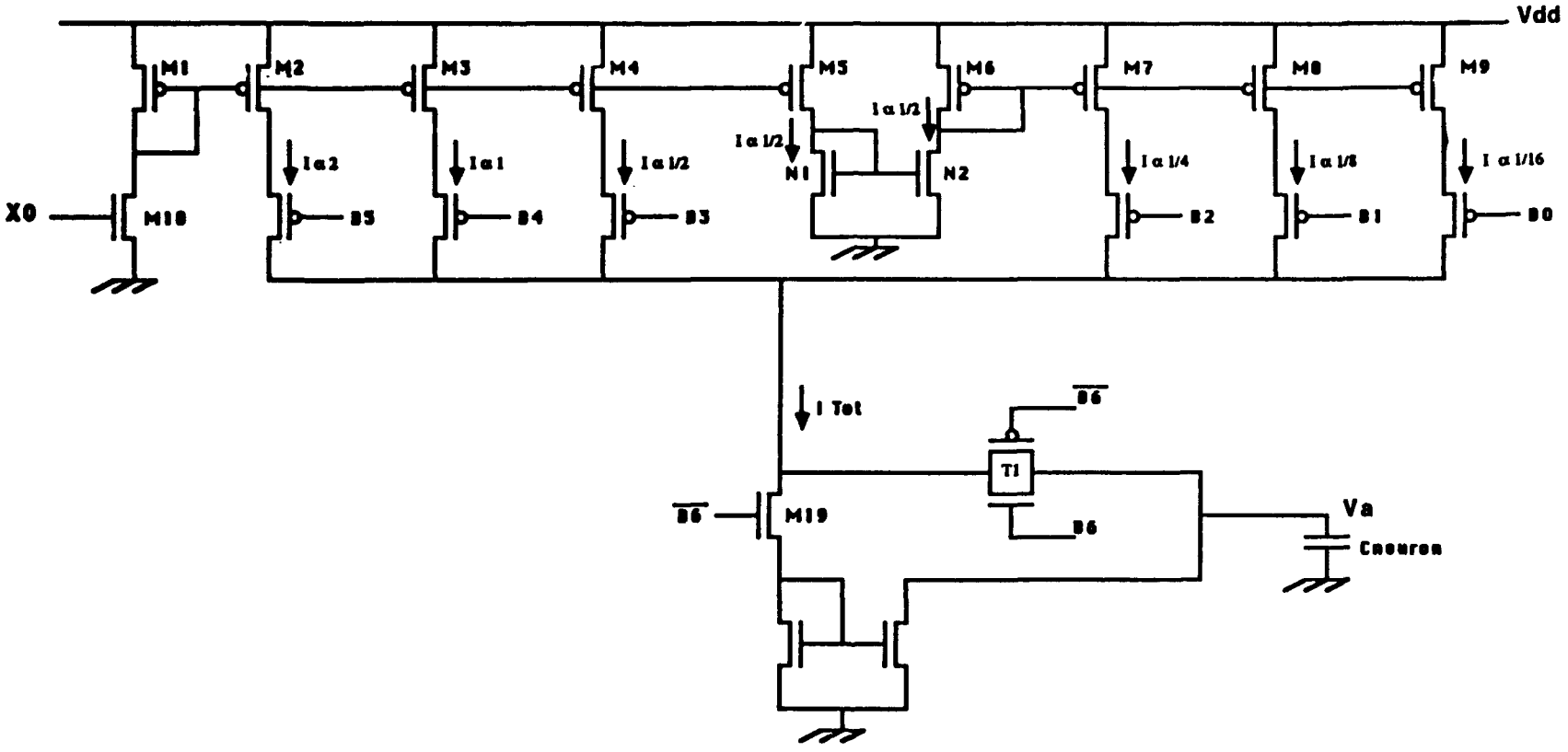


Fig 8: Seven bit (6 bit + sign) current division synapse circuit

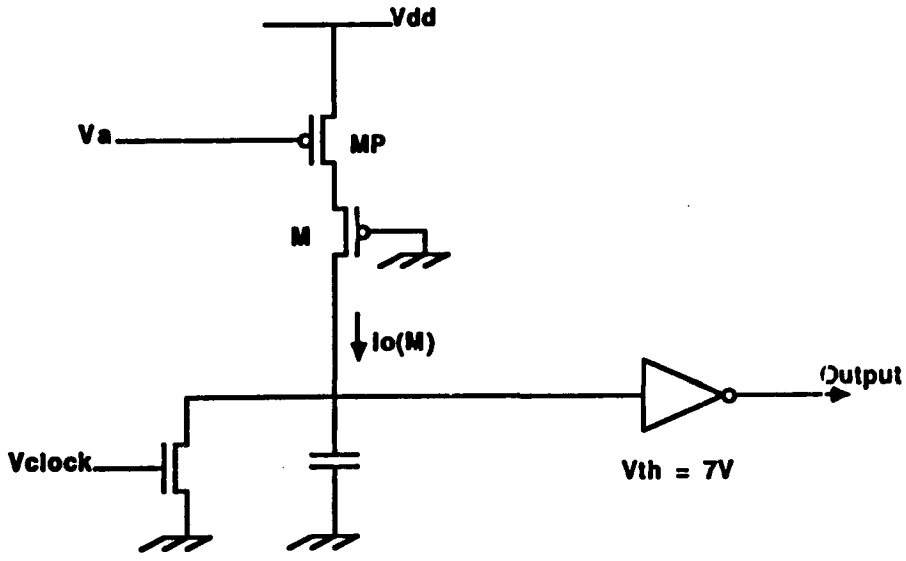


Fig 9: Pulse-width modulating neuron circuit

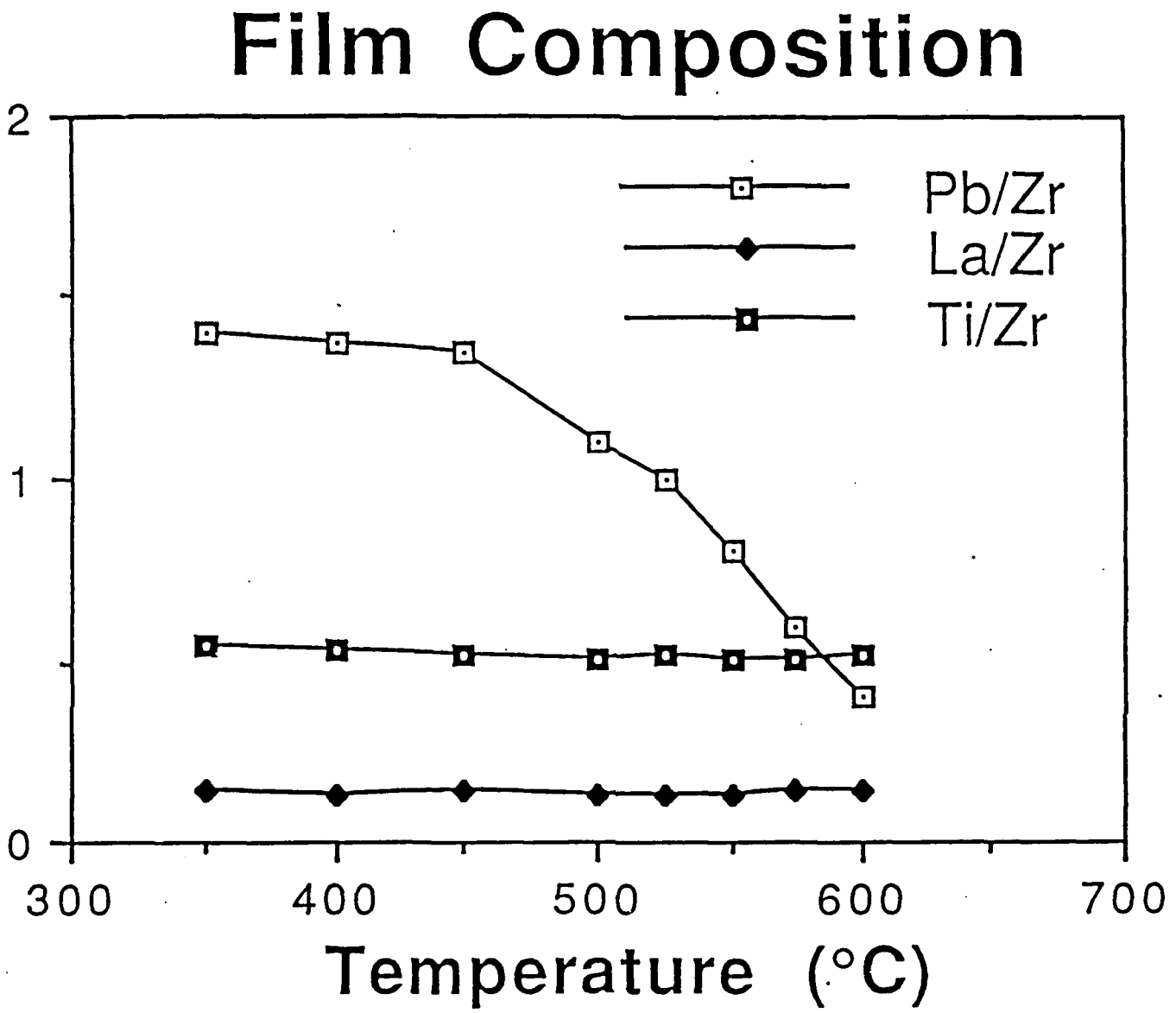


Figure 10. The plot of deposited PLZT film composition vs. substrate temperature.

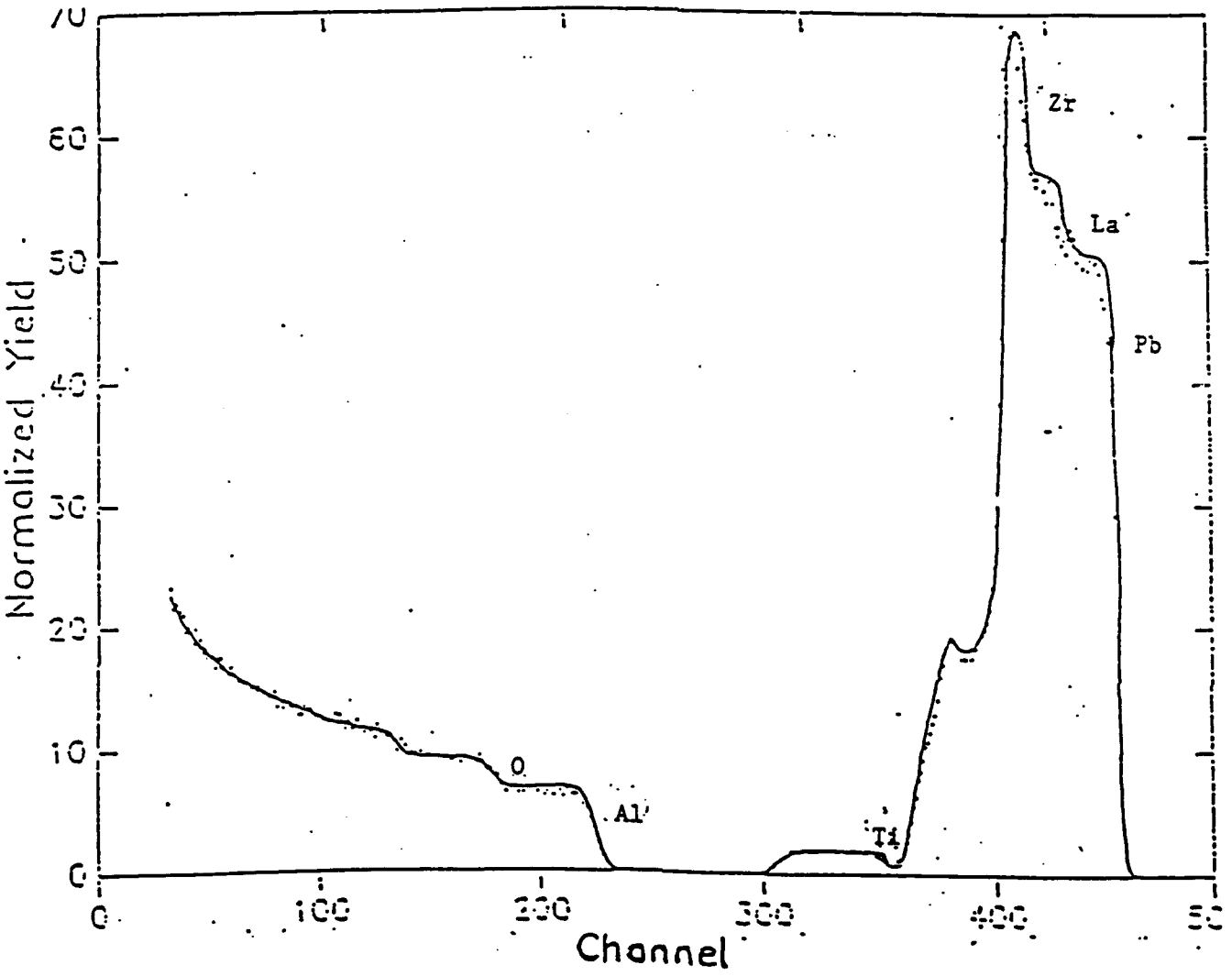


Figure 11. RBS profile of a 9/65/35 PLZT film of 0.45 μm thickness deposited on sapphire. Solid curve is simulation result while dotted curve is measurement result.

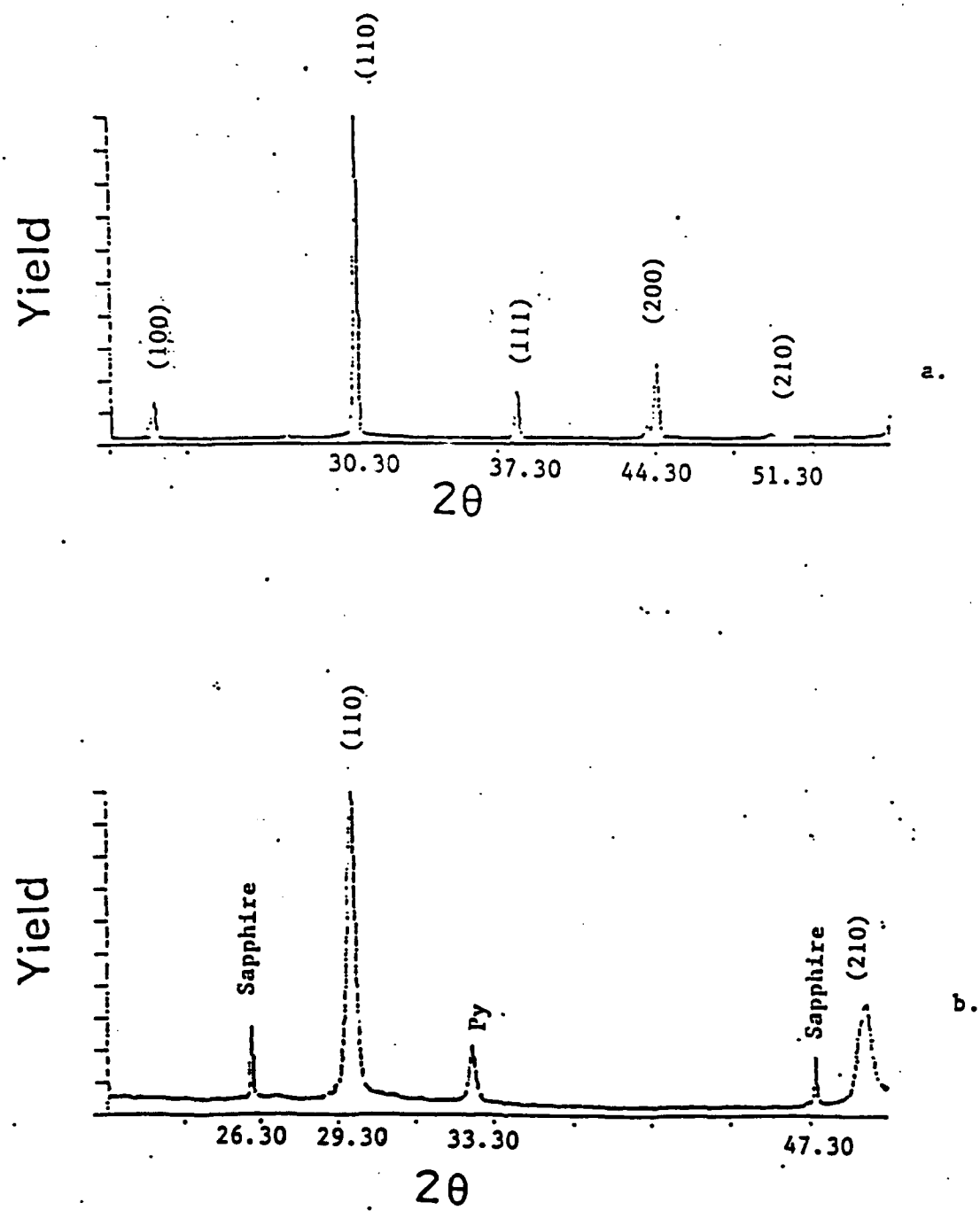


Figure 12. X-ray diffraction spectra of a) Bulk 9/65/35 PLZT, b) Thin 9/65/35 PLZT film deposited on sapphire.

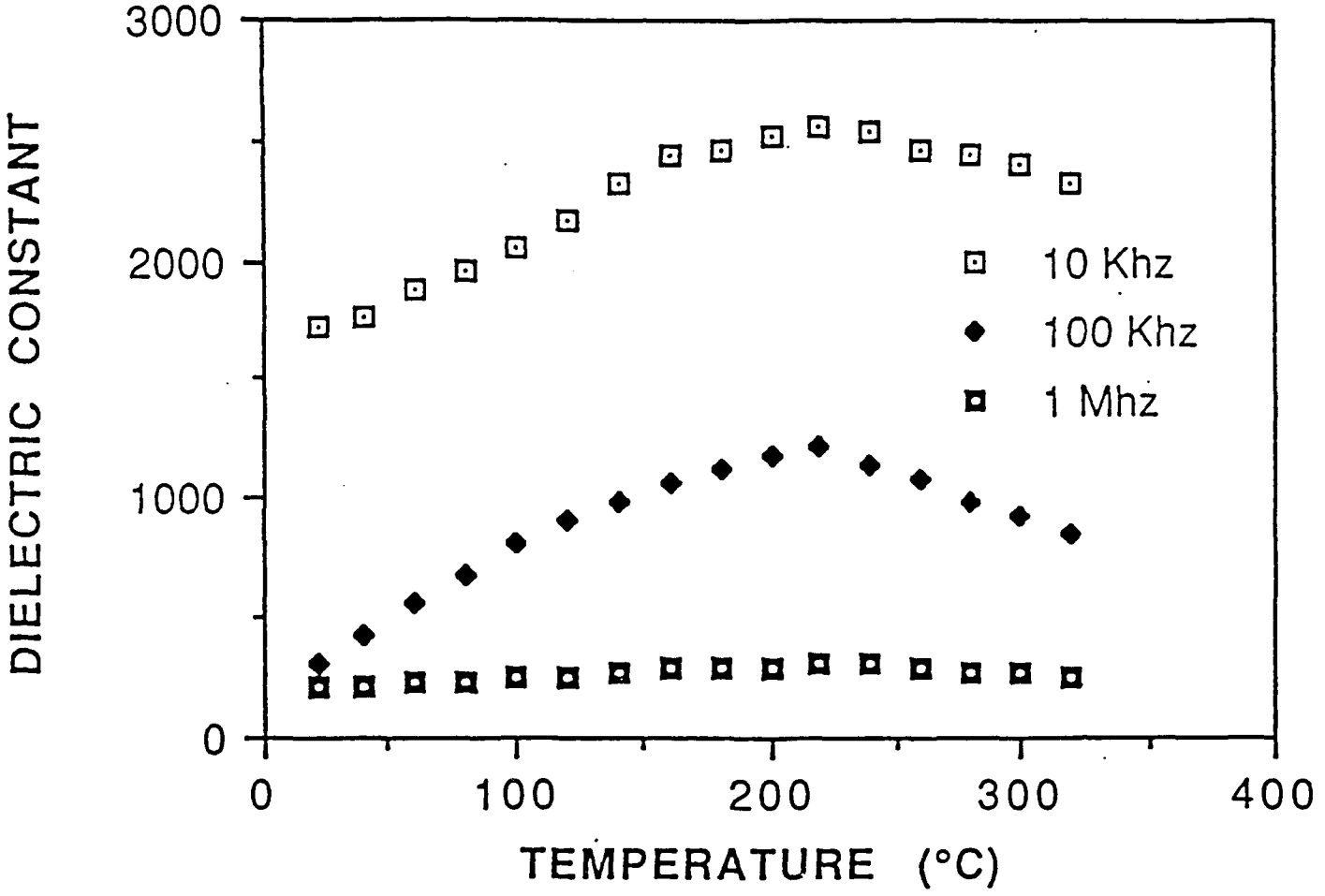


Figure 13. Variation of the relative dielectric constant of deposited 9/65/35 PLZT film with temperature.

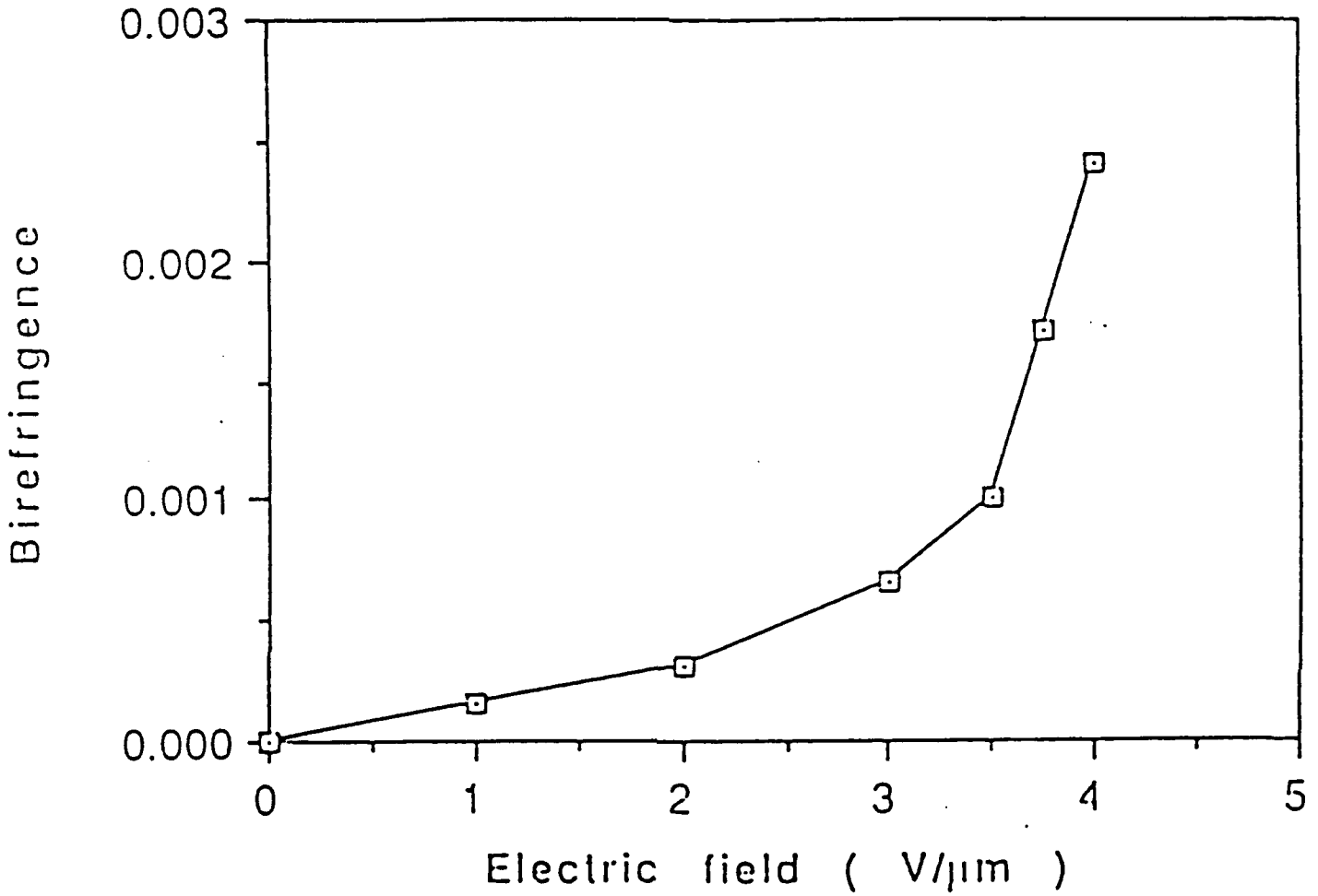


Figure 14. Change of birefringence vs. applied electric field in deposited 9/65/35 PLZT film. Quadratic electro-optic effect is $R = 0.6 \times 10^{-16} m^2/V^2$ (for bulk PLZT, $R = 3.6 \times 10^{-16} m^2/V^2$).

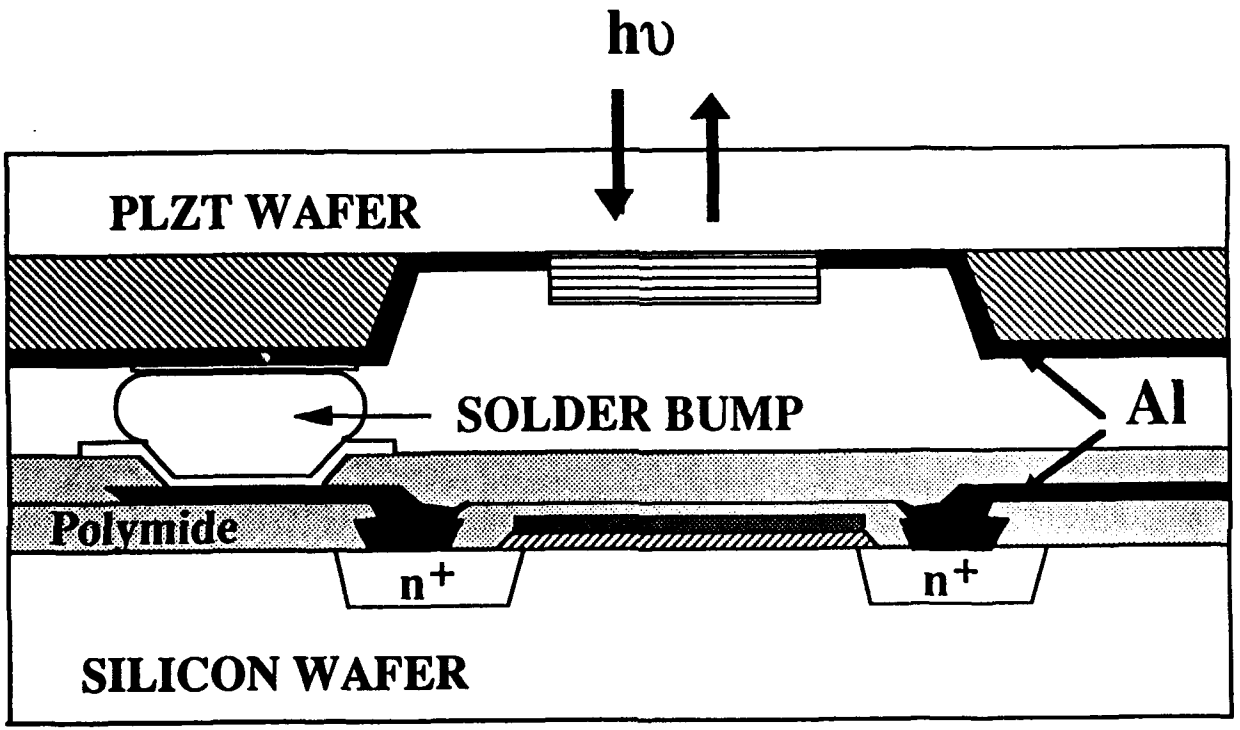
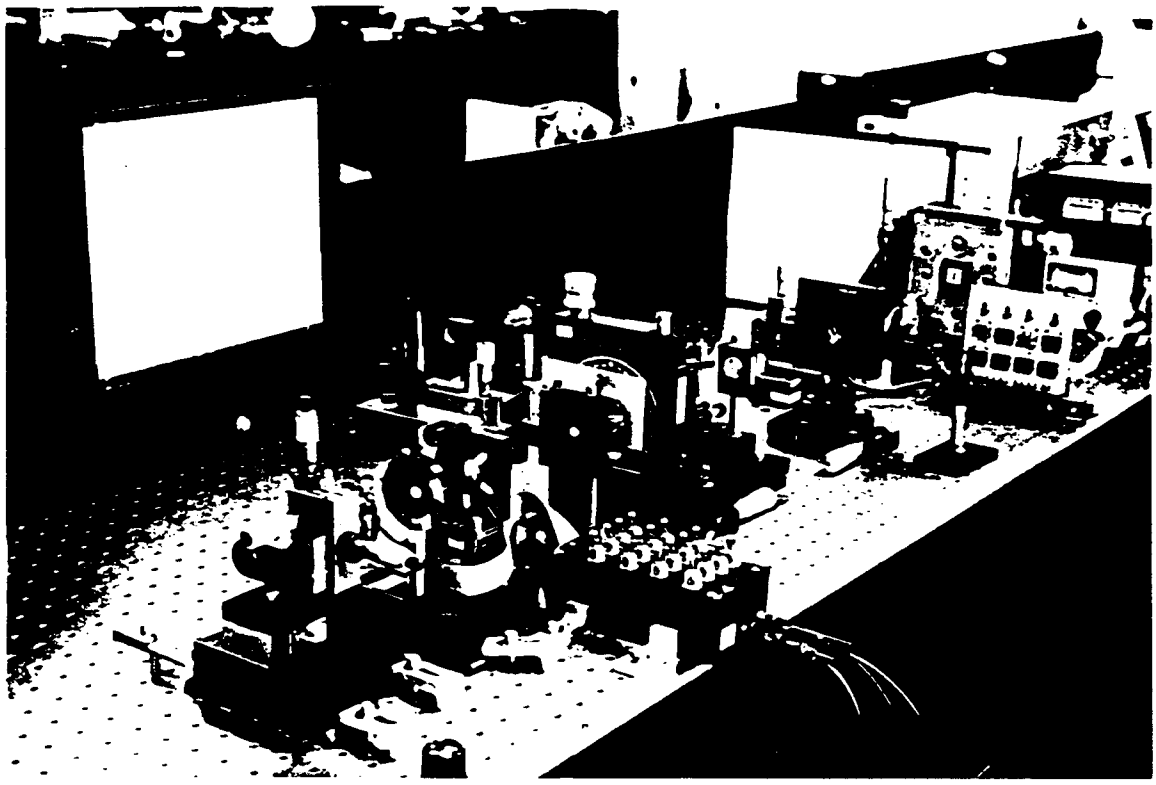
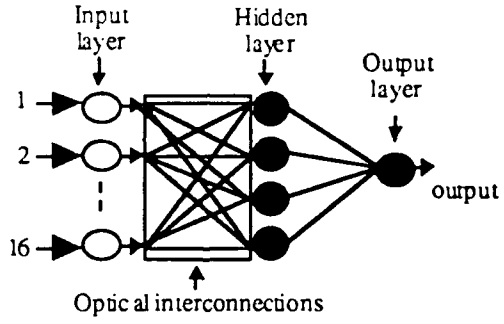


Fig 15: Flip-chip Bonded PLZT on Silicon



(16 a)



(16 b)

Fig. 16 : (a) Picture of the OE neural system showing: 1. Input polarizer 2. Filtering/collimation optics 3.Lenslet array / PLZT spatial light modulator (input layer) 4. Demagnification optics 5. Polarizing beam splitter 6. Replication lens/hologram 7. OE neural chip (hidden layer) 8. Output display / output layer
(b) Its network equivalent

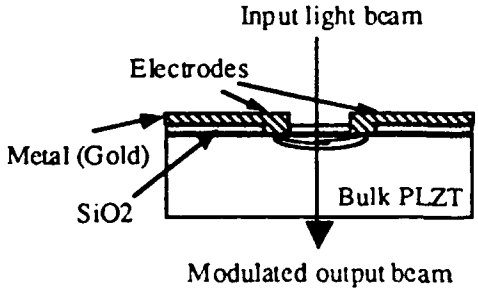


Fig 17 : Cross section of an electrically addressed PLZT light modulator

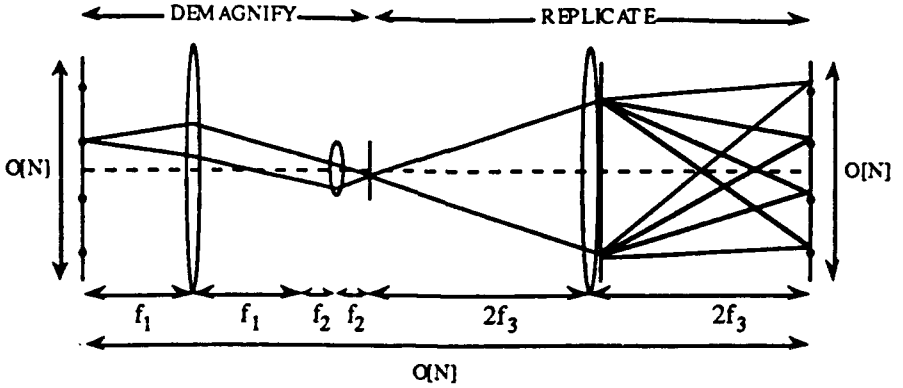


Fig.18: The DSTOP optical system

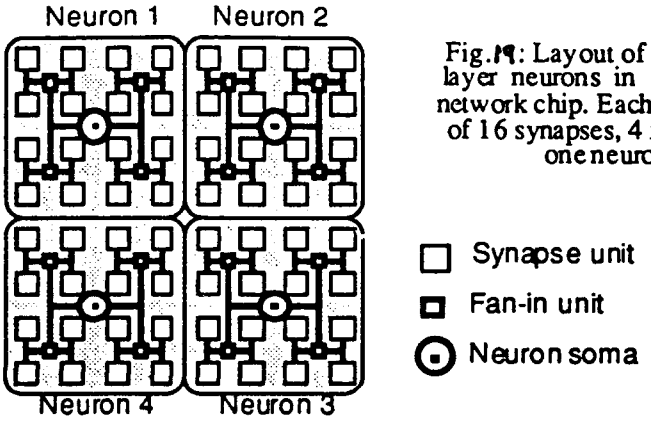


Fig.19: Layout of the four hidden layer neurons in the OE neural network chip. Each neuron consists of 16 synapses, 4 fan-in units and one neuron soma

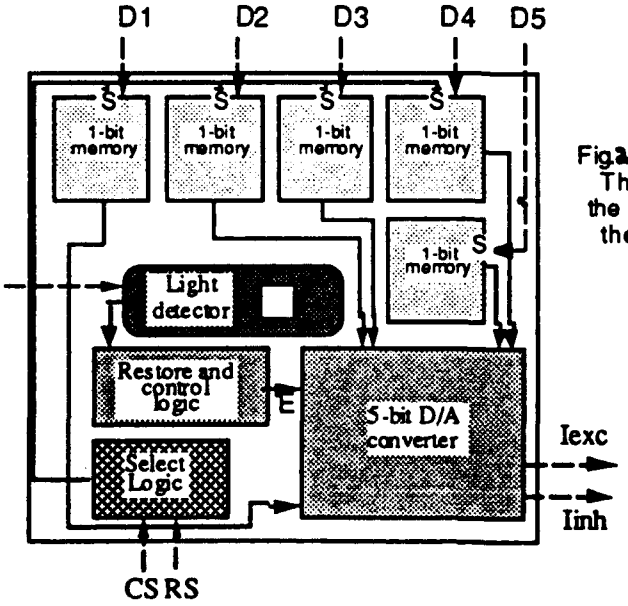
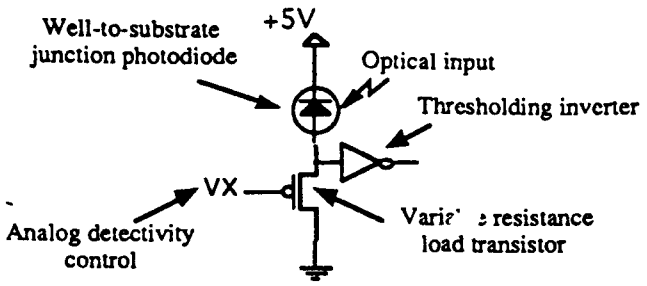


Fig.20: Functional diagram of the synapse unit. The synapse uses a light detector to detect the optical input, 5-bit digital memory to store the synaptic weight and a D/A converter to generate an analog output signal



Photodiode area : 15×15μm2
Responsivity : 0.3 A/W
Speed : 10MHz/100μW

Fig.21 : Design and characteristics of the photodetector circuit

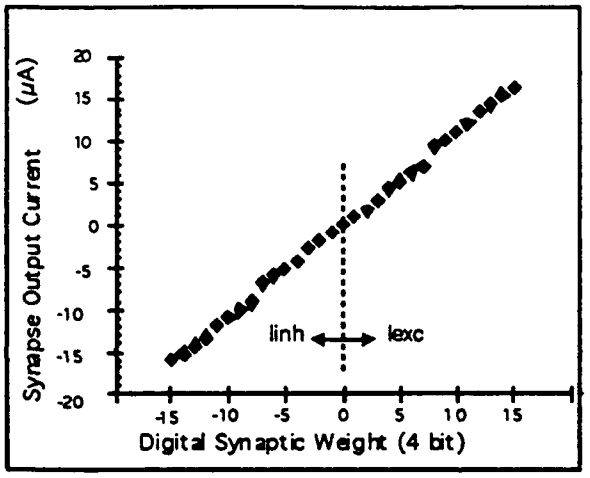


Fig.22: Synapse output current as a function of the stored digital synaptic weight

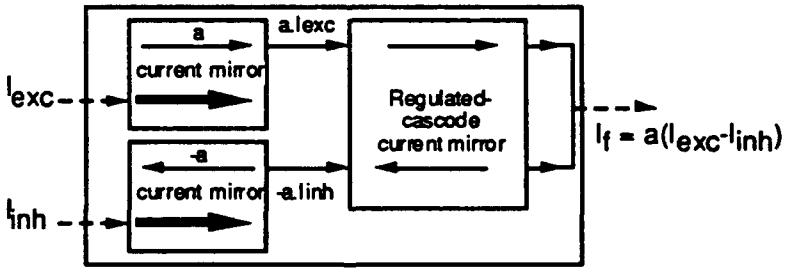


Fig.23 :Functional diagram of the fan-in(dendrite) unit

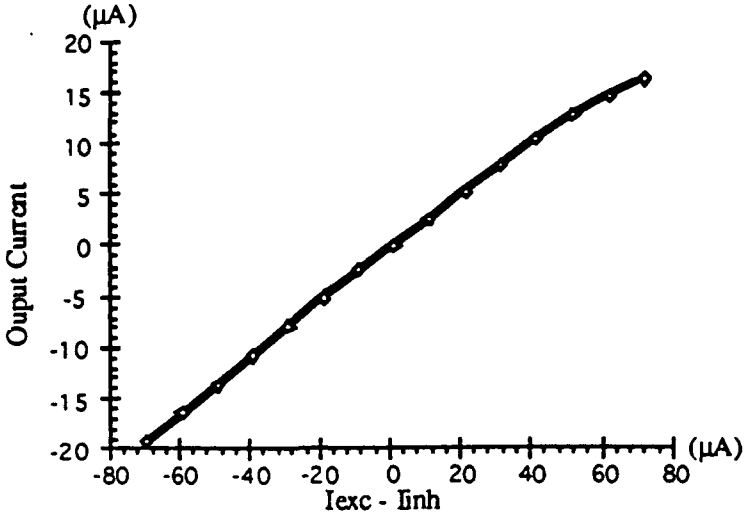


Fig 24 : Fan-in unit transfer function

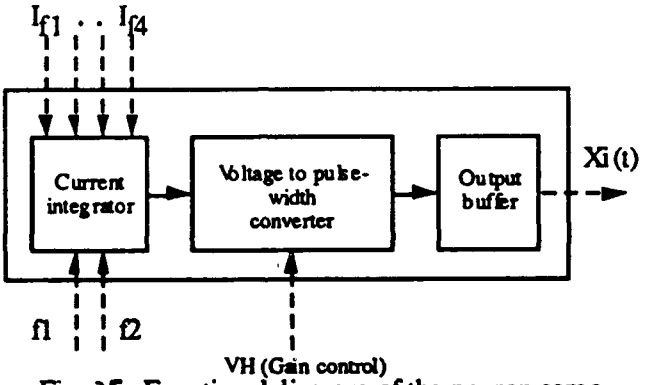


Fig. 25: Functional diagram of the neuron soma

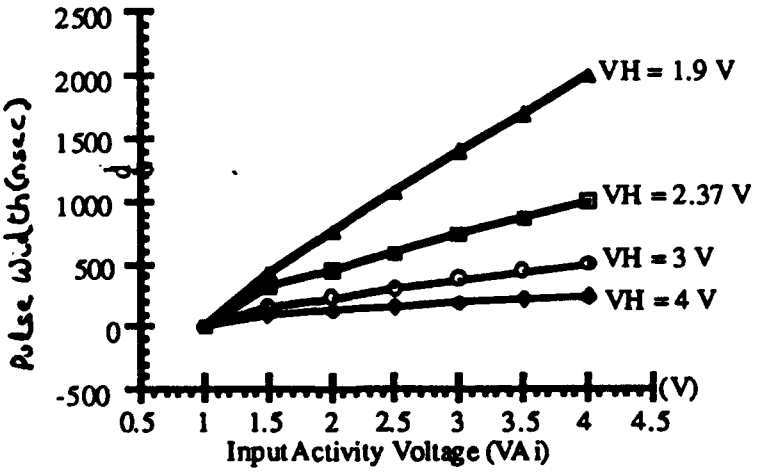


Fig. 26: Neuron transfer function with defferent gains

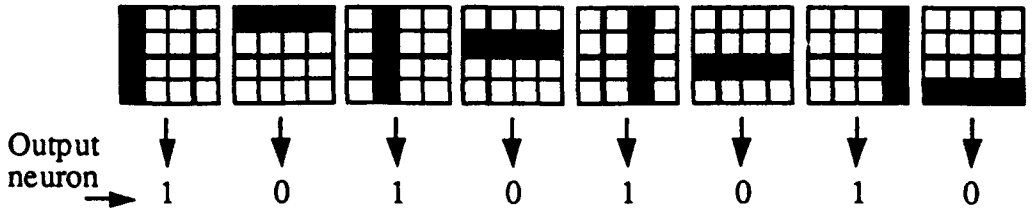


Fig. 27 : Test patterns applied to the system together with the corresponding output neuron states